DOCUMENT RESUME

ED 171 801 TM 009 508

AUTHOR Brown, F. G.; Moss, Jacque
TITLE Empirical Methods for Studying Sex Bias in Achievement Test Items.
PUB DATE 9 Apr 79
NOTE 14p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; College Students; *Content Analysis; Females; Higher Education; Item Analysis; Males; *Performance Factors; Reading Tests; *Scores; *Sex Discrimination; *Test Bias; *Test Items

ABSTRACT

Two studies were conducted of the effects of sex biased items on achievement test scores obtained by male and female college students. Sex bias was defined as a disproportionate number of references to one sex. Study I used 1,008 classroom test items. Items were classified as male sex referent, female sex referent, no sex referent (neutral), or ambiguous sex referent (other). Data showed a small, but statistically significant, sex bias effect, whereby students performed better on items containing referents to their own sex than on other classes of items. Study II was experimental and controlled for sex bias by using two forms of a 3,000-word passage on defense mechanisms and two tests: one with predominantly male sex referents and the other with predominantly female sex referents. Half of the students took a test having the same predominant sex referents as the passage they had read; the other half took a test with the opposite sex referents predominating. There were no statistically significant effects. The results of the two studies suggest that sex bias in test items has in actuality, minimal effects on test scores. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

F. G. Brown

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM"

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED171801

TM009 508

EMPIRICAL METHODS FOR STUDYING SEX BIAS IN ACHIEVEMENT TEST ITEMS

F. G. Brown, Iowa State University

Jacque Moss, University of Iowa

Abstract. Two studies of the effects of sex bias on achievement test performance were conducted. Sex bias was defined as a disproportionate number of references to one sex. Both studies investigated the question: What effect does inclusion of sex biased items have on the test scores obtained by males and females? This question can be answered by looking at item by examinee interactions.

Study I used 1008 items from regularly scheduled classroom exams. Items were classified as having a male sex referent, a female sex referent, no sex referent (neutral), or an ambiguous sex referent (other). The data showed a small, but statistically significant, sex bias effect. That is, students performed better on items containing referents to their own sex than on other classes of items.

Study II was an experimental study which controlled for sex bias in both the learning task and the test items. Students read a 3000-word passage, then took an achievement test over this material. Two forms of the passage and test were used; one with predominantly male sex referents and the other with predominantly female sex referents. Half of the students took a test having the same predominant sex referents as the passage they had read; the other half took a test with the opposite sex referents predominating. There were no statistically significant effects.

Taken together, the results of the two studies suggest that sex bias in test items has minimal effects on test scores. Thus, arguments for reducing sex bias on achievement test items should be made on other bases than their effect on test performance.

Presented at: National Council on Measurement in Education
1979 Annual Convention (San Francisco)
April 9, 1979

Currently, an issue of great concern to test developers and users is the possible effects of bias in educational and psychological tests. These concerns have generally focused on either cultural/ethnic/racial or sex bias. This paper will be concerned with the latter category - sex bias. However, some of the issues discussed will also have relevance to other types of bias in tests.

The issues relating to test bias involve a wide variety of legal, ethical, political, and psychometric problems and considerations. Needless to say, there is much disagreement over the nature and extent of test bias. There is even disagreement over what is meant by bias or, conversely, fairness. As the Standards for educational and psychological tests states:

> It is important to recognize that there are different definitions of fairness, and whether a given procedure is or is not fair may depend on the definition accepted. (Standards, 1974, p. 44)

The dictionary defines bias in terms of partiality or prejudice. Using this definition, a testing procedure would be biased if it showed partiality towards or prejudice against certain individuals or groups of individuals. But, implicit in this definition is the assumption that the partiality or prejudice is the result of consideration of irrelevant factors. Thus, more specifically, a testing procedure would be biased if consideration of factors irrelevant to the purposes of the testing resulted in different inferences or implications being drawn from the scores of different individuals or subgroups of a population.

Sex bias. When discussing sex bias we have to distinguish between two concepts: sex stereotyping and sex bias. By "sex stereotyping" we mean the use of conventional, rigid, and/or oversimplified presentations of males and females - e.g., picturing males as strong and intelligent or females

as weak, emotional, and dependent. By "sex bias" we mean use of a disproportionate number of references to members of one sex, even when the content of the material does not require such emphases - e.g., the disproportionate use of masculine pronouns in textbooks. Although these two concepts are related, they can be separated both logically and empirically.

Another important consideration in studying test bias is where and how the bias operates. Both Diamond (1977) and Reschley (1979) have described various types of bias. Content bias occurs when the content and/or phrasing of the test items favors one group over another. Atmosphere bias occurs when the testing conditions (e.g., motivation, prior testing experience, examiner-examinee interaction) favor one group. The third type of bias is bias in use. Here Reschley differentiates between two types of bias: bias in prediction occurs when the use of a particular prediction equation or strategy is unfair to some subgroup while bias in terms of social consequenses occurs when different treatments result from test use. In this paper we will be concerned with content bias.

There is ample evidence that sex bias (as defined above) is present in textbooks and other reading material (see, e.g., Frasher & Walker, 1972; Grambs, 1972). There is also evidence that sex bias occurs on standardized achievement tests (Faggen-Steckler, McCarthy & Tittle, 1974; Tittle, 1974). Several authors (Diamond 1976, 1977; Tittle, 1975) have suggested ways to study and control sex bias in tests. What has not been investigated, however, is how much effect, if any, sex bias has on performance. As Faggen-Steckler et al state:

> Little, if anything, is known of the developmental consequenses of the existing bias in school materials. To date there has been no significant study which demonstrated affective or behavioral changes in school-age children due to the constant verbal and written barrage

> of language sex inbalance and sex-role stereotyping to which, it now seems evident, they are exposed. (Faggen-Steckler et al, 1974, p. 159)

To the authors' knowledge, no such studies have appeared since that time - a surprising fact considering the interest in test bias.

The present studies. Our concern in the present studies was with the effects of sex bias in test items on performance. More specifically, we were interested in what effect inclusion of disproportionate numbers of referents to one sex in test items would have on the scores obtained by males and females. Thus, we were looking for item by examinee interactions. The presence of such interactions would have obvious implications for the construction, interpretation, and use of (achievement) tests.

As noted above, there is little empirical evidence regarding this question. One study (Donlon, 1971) reported that mean differences on the mathematics section of the Scholastic Aptitude Test might be partially reduced by changes in test items. Related evidence from studies of race by item interactions is contradictory. For example, Medley & Quirk (1974) found evidence of such interactions using items from the National Teacher Examination while Schmeiser & Ferguson (1978) found no such interactions when using items on the ACT Assessment.

The purpose of the studies described below can be viewed in either of two ways. First, they illustrate two methods that can be used to study the effects of sex bias and, second, they provide some evidence regarding the magnitude of such effects.

Study I: Bias on classroom exams

The question addressed in this study was: Given a particular (nonbiased) content domain, will students perform better on test items that contain

references to their own sex than on items containing references to the opposite sex or phrased in neutral terms?

Method. Data were collected on 1008 items from regularly scheduled exams in an educational psychology course. By using an educational psychology course the effects of sex stereotyping were minimized as the items were phrased in terms of instructional situations and the test takers (college students) had had extensive experience with both male and female teachers and male and female students.

Items were classified in terms of the sex-referents contained in the items. An example[1] of a male item was:

> Johnny is constantly talking in class. The teacher tries to change his behavior by ignoring him when he talks out of turn. This technique is called: A) discrimination B) extinction C) negative reinforcement D) satiation.

An example of a female item was:

> A class of 100 students takes a 40 item test. Jane gets 28 items correct and scores higher than 60 students. Her percentile rank is: A) 70 B) 60 C) 40 D) 30.

Items with no sex-referents or containing a name with sex not indicated (e.g., Piaget) were classified as "neutral". Items with both male and female referents or ambiguous referents (e.g., the teacher, a student) were classified as "other". The basic data unit was the item difficulty index (percent correct) for males and females on each item. The average number of students in the item analyses was 40 males (range 26 - 55) and 100 females (range 81 - 123).

Results. Performance on the 1008 items was classified into four content categories (M-referent, F-referent, neutral, other) and three performance categories (M>F, F>M, M = F). Chi-square analysis showed that the

---

1. Emphasis in examples only; it was not present on the actual test.

distributions were significantly different at the .01 level (see Table 1).

Analysis of mean differences (F - M) showed that, across all item categories, females averaged 3.1% more correct answers per item than males. When items contained a female-referent the difference was larger (4.6%) and when the item contained a male-referent the difference was smaller (1.2%). Thus the direction of the effect was for students to perform relatively better on items containing a referent to their own sex.

One limiting factor in the study was a possible confounding of item content with sex-referents. That is, the differential performance of males and females on (some) items may have been due to item content rather than to sex bias. As no attempt was made to control this factor, no direct test of this hypothesis was possible. Some interesting questions were raised, however, when looking at items exhibiting large sex differences. An informal content analysis was made using a 15% difference in difficulty between males and females as the criterion. The items on which males performed better often involved quantitative reasoning (e.g., converting standard scores to percentiles, interpreting the graph of an interaction). The items on which females performed better involved motivational concepts, behavioral objectives, knowing differences between types of learning and instructional methods, defining terms, and drawing conclusions from studies. These results suggest that differences in content and/or cognitive processing required may be important, regardless of the sex bias in an item. (Cf. Maccoby & Jacklin, 1974)

Conclusions. The results indicate that sex-referents in test items have a significant, albeit small, effect on exam performance. Thus, if an instructor includes a disproportionate number of items with references to one sex and/or when scores are summed over several tests (as when

Table 1. Performance of males and females on various classes of items.

	Frequencies			Proportions			
Item type	F>M	F=M	M>F	F>M	F=M	M>F	Mean Diff*
Neutral	442	30	209	.649	.044	.307	.034
Male	95	14	80	.503	.074	.413	.012
Female	47	7	22	.618	.092	.289	.046
Other	45	3	14	.726	.048	.226	.038
TOTAL	629	54	325	.622	.054	.322	.031

$\chi^2 = 19.48$ df = 6 $p < .01$

*Mean Diff = Mean difference = difference in percent correct responses to an item (F - M).

determining grades), an advantage may be given to students of that sex. Therefore, instructors should attempt to balance the number of sex-referents in their test items.

Study II: A simulated learning experience

One factor that was not controlled in the first study was the content of the reading material on which the test items were based. Particularly, there was no control for possible sex bias in the reading material. As characteristics of the reading material influence retention (McConkie, 1978), it would be desireable to control for possible bias in the reading material. Thus a second study was designed which controlled for the effects of sex-referents in both the reading passages and the test items.

Method. A 3000-word reading passage was constructed. The topic was defense mechanisms, a topic that subjects had not yet covered in their courses and which would be equally applicable to males and females. Two parallel forms of the passage were written, one masculine and one feminine. There were no content differences in the passages; they differed only in noun and pronoun references. In the male passage, 60% of the sex-referents were to males, 20% to females, and 20% were neutral. In the female passage the percentages were reversed. There were a total of 110 sex-referents in the passage. The opposite and neutral referents were included to make the manipulation less obvious. A random half of the subjects read each passage.

After they read the passage, students answered 25 multiple-choice items covering the material in the passage. Sex-referents in the test items were distributed in the same proportions as in the reading passage (60%, 20%, 20%). Half of the subjects took a test having the same predominant sex-referents

as the reading passage and half took a test having the opposite referents predominant. Again assignment was random.

The subjects were 120 volunteers from introductory psychology courses. Four scores were obtained from each test: total (all 25 items), predominant-sex items (5), opposite-sex items (5), and neutral items (5). As the results of the analyses for each type of score were similar, only the results for the total (composite) score will be reported.

Results. The analysis was a 2 x 2 x 2 analysis of variance with the between subjects variables being sex of student, predominant sex-referents in the reading passage, and predominant sex-referents in the test items. There were no significant F's, either for main effects or interactions. Thus, when bias in the learning material (reading passage) was controlled, there were no differences in performance as a function of bias in the test items.

Although the analysis of variance showed no significant results, there were some interesting trends in the data (see Table 2). When experimental groups were rank ordered in terms of performance on the criterion test, there was a sex of subject by sex of passage match (i.e., males reading male passages, females reading female passages) in three of the top four groups but in only one of the four lowest groups. Similar results were found using subject by item and passage by item matches. The largest difference (2.8 points) was between men who read a female passage and took a female test (mean = 17.13) and women in the same experimental condition (i.e., female passage, female test; mean = 19.93). This would again suggest that use of disproportionate numbers of sex-referents may have an effect on test scores.

Table 2. Mean composite test scores by experimental treatment group.

A. Individual Treatment Conditions

Subjects	M	M	M	M	F	F	F	F
Passage	M	M	F	F	M	M	F	F
Items	M	F	M	F	M	F	M	F
Test score	18.80	19.07	18.20	17.13	17.40	19.80	18.13	19.93

B. Interactions

Subjects x Passage

M subjects, M passage = 18.94 M subjects, F passage = 17.66

F subjects, M passage = 18.60 F subjects, F passage = 19.03

Subjects x Items

M subjects, M items = 18.50 M subjects, F items = 18.10

F subjects, M items = 17.76 F subjects, F items = 19.86

Passage x Items

M passage, M items = 18.10 M passage, F items = 19.44

F passage, M items = 18.16 F passage, F items = 18.53

C. Composite means

Subjects	M = 18.30	F = 18.81
Passage	M = 18.77	F = 18.35
Items	M = 18.13	F = 18.98

Conclusions. The results of the analysis of variance indicated that when sex-referents are controlled in the reading materials, there are no significant differences as a function of bias in the test items. However, further analyses showed a trend in the same direction as found in Study I - students performed better on items (and reading material) that contains references to their own sex.

Discussion

The results of the two studies were not entirely consistent. In the first study, which controlled only for bias in test items, there were significant, though small, sex bias effects. In the second study, where bias in both the learning materials and test items was controlled, there were no significant differences in test performance.

While the differences were small (and were statistically insignificant in the second study) the data were consistent in showing that a small advantage may accrue if the test items (and/or reading materials) are phrased using references to the same sex as the test taker. This effect will be more pronounced when the disproportionality of the sex-referenced items is high and/or when scores are cumulated over several tests. While the effects of sex bias on any single test will be small, when cumulated over a course, or several years of education, the effects could indeed become significant. Thus, test constructors should attempt to balance sex references on their tests.

Two further points should be made. First, even though use of a disproportionate number of references to one sex appears to have minimal effects on test performance, it may contribute to sex stereotyping. If

this is the case, sex bias should be avoided. On the other hand, because the magnitude of the effects on performance was small, arguments against sex bias (and sex stereotyping) in educational materials, including tests, should be based on stronger arguments than their effect on performance.

Second, most of the discussion in the literature on sex bias is based on conjecture and not buttressed by empirical data. The typical argument states that such-and-such may occur but do data is presented to show that such-and-such does, in fact, occur. The two studies described above illustrate how the effects of sex bias can be studied empirically.

REFERENCES

Diamond, E. E. Minimizing sex bias in testing. Measurement and Evaluation in Guidance, 1976, 9, 28-34.

Diamond, E. E. Testing: The baby and the bath water are still with us. Pp. 31-44 in Testing and the public interest (Proceedings of the 1976 ETS Invitational Conference). Princeton, N.J.: Educational Testing Service, 1977.

Donlon, T. F. Content factors in sex differences on test questions. Paper presented at the meeting of the New England Educational Research Organization, 1971. [Cited by C. K. Tittle. Fairness in educational achievement testing. Education and Urban Society, 1975, 8(1), 86-103.]

Faggen-Steckler, J., McCarthy, K. A., & Tittle, C. K. A quantitative method for measuring sex "bias" in standardized tests. Journal of Educational Measurement, 1974, 11, 151-161.

Frasher, R. & Walker, A. Sex roles in early reading textbooks. The Reading Teacher, 1972, 25, 741-749.

Grambs, J. D. Sex-stereotypes in instructional materials, literature, and language: A survey of research. Women Studies Abstracts, 1972, 1, 1-4 & 91-94.

Maccoby, E. E. & Jacklin, C. N. The psychology of sex differences. Stanford, Ca.: Stanford University Press, 1974.

McConkie, G. W. Learning from text. Pp. 3-48 in L. S. Shulman (Ed.). Review of research in education 5. Itasca, Ill.: Peacock, 1978.

Medley, D. M. & Quirk, T. J. The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examination. Journal of Educational Measurement, 1974, 11, 235-245.

Reschley, D. J. Nonbiased assessment. In G. Phye & D. Reschley (Eds.). School psychology: Perspectives and issues. New York: Academic Press, 1979.

Schmeiser, C. B. & Ferguson, R. L. Performance of black and white students on test materials containing content based on black and white cultures. Journal of Educational Measurement, 1978, 15, 193-200.

Standards for educational and psychological tests. Washington, D.C.: American Psychological Association, 1974.

Tittle, C. K. Sex bias in educational measurement: Fact or fiction? Measurement and Evaluation in Guidance, 1974, 6, 219-226.

Tittle, C. K. Fairness in educational achievement testing. Education and Urban Society, 1975, 8(1), 86-103.